Open-reading-frame sequence tags (OSTs) support the existence of at least 17,300 genes in *C. elegans*

Jérôme Reboul[1]*, Philippe Vaglio[1]*, Nia Tzellas[1], Nicolas Thierry-Mieg[1,2], Troy Moore[3], Cindy Jackson[3], Tadasu Shin-i[4], Yuji Kohara[4], Danielle Thierry-Mieg[5], Jean Thierry-Mieg[5], Hongmei Lee[6], Joseph Hitti[6], Lynn Doucette-Stamm[6], James L. Hartley[7], Gary F. Temple[7], Michael A. Brasch[7], Jean Vandenhaute[8], Philippe E. Lamesch[1,8], David E. Hill[1] & Marc Vidal[1]

*These authors contributed equally to this work.

The genome sequences of *Caenorhabditis elegans*, *Drosophila melanogaster* and *Arabidopsis thaliana* have been predicted to contain 19,000, 13,600 and 25,500 genes, respectively[1–3]. Before this information can be fully used for evolutionary and functional studies, several issues need to be addressed. First, the gene number estimates obtained *in silico* and not yet supported by any experimental data need to be verified. For example, it seems biologically paradoxical that *C. elegans* would have 50% more genes than *Drosophilia*. Second, intron/exon predictions need to be tested experimentally. Third, complete sets of open reading frames (ORFs), or "ORFeomes,"[4] need to be cloned into various expression vectors. To address these issues simultaneously, we have designed and applied to *C. elegans* the following strategy. Predicted ORFs are amplified by PCR from a highly representative cDNA library[4] using ORF-specific primers, cloned by Gateway recombination cloning[4–6] and then sequenced to generate ORF sequence tags (OSTs) as a way to verify identity and splicing. In a sample (n=1,222) of the nearly 10,000 genes predicted *ab initio* (that is, for which no expressed sequence tag (EST) is available so far), at least 70% were verified by OSTs. We also observed that 27% of these experimentally confirmed genes have a structure different from that predicted by GeneFinder. We now have experimental evidence that supports the existence of at least 17,300 genes in *C. elegans*. Hence we suggest that gene counts based primarily on ESTs may underestimate the number of genes in human and in other organisms.

The *C. elegans* Sequencing Consortium (http://www.wormbase.org) predicts 18,959 genes (in version WS9) that can be classified according to the availability of data confirming their existence (Fig. 1; see our working definition of 'genes' in Methods). *C. elegans* researchers collectively have cloned and sequenced 784 genes (the 'community' genes), whereas the *C. elegans* EST project has identified 9,356 genes (the 'EST' genes; Y.K. *et al.*, manuscript in preparation). Because 637 genes correspond to both community and EST genes, this brings the total number of *C. elegans* genes whose existence has been confirmed experimentally to 9,503 (Fig. 1, left). Of these genes, 796 are not predicted by GeneFinder, which illustrates the rate of false negatives of this genome annotation tool (Fig. 1, right). This leaves 8,707 experimentally confirmed genes that are predicted by GeneFinder. It should be noted that those 8,707 genes correspond to 9,071 predicted genes (Fig. 1). This is due to a slightly higher tendency of GeneFinder to predict two genes for one, rather than one gene for two (referred to as 'split' or 'joined' genes, respectively).

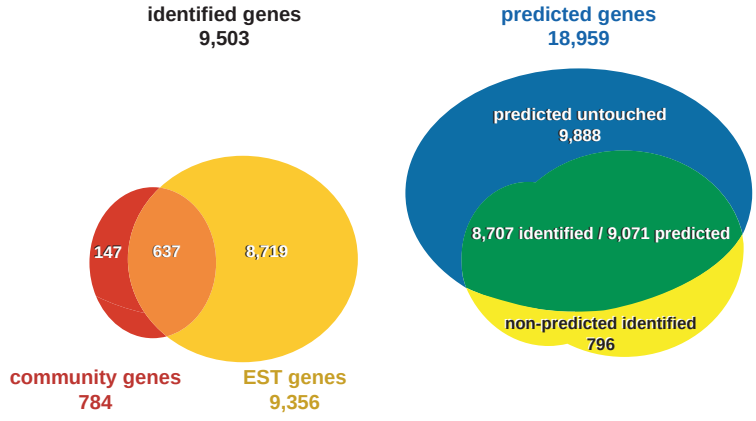


Fig. 1 Current compilation of protein-encoding genes in *C. elegans*. The left-hand VENN diagram shows the 2 classes of the 9,503 experimentally identified genes. Of these genes, there are 784 community genes obtained from traditional cloning (red circle) and 9,356 EST genes identified on the basis of expressed sequence tags (yellow circle). Of 784 community genes, 147 have not been identified by EST analysis. The right-hand VENN diagram shows the intersection between the total number of predicted genes (18,959; blue oval) and the experimentally identified genes (yellow circles). The green region of intersection shows the 8,707 experimentally identified genes that have been predicted. The blue region of non-intersection shows the 9,888 predicted genes that have not been experimentally confirmed until now. The yellow region of non-intersection shows the 796 identified genes that are not predicted by GeneFinder. In addition, in some cases, pairs of seemingly separate predicted genes correspond to a single experimentally identified gene. Conversely, single predictions sometimes correspond to pairs of experimentally identified genes. The last two facts explain why 8,707 genes correspond to 9,071 predicted genes.

[1]*Dana-Farber Cancer Institute and Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA.* [2]*Laboratoire LSR-IMAG, St-Martin D'Heres, France.* [3]*Research Genetics, Huntsville, Alabama, USA.* [4]*Genome Biology Laboratory, National Institute of Genetics, Mishima, Japan.* [5]*National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA.* [6]*Genome Therapeutics Corp., Waltham, Massachusetts, USA.* [7]*Life Technologies Inc., Rockville, Maryland, USA.* [8]*Département de Biologie, Facultés Universitaires Notre-Dame de la Paix, Namur, Belgium. Correspondence should be addressed to M.V. (marc_vidal@dfci.harvard.edu).*

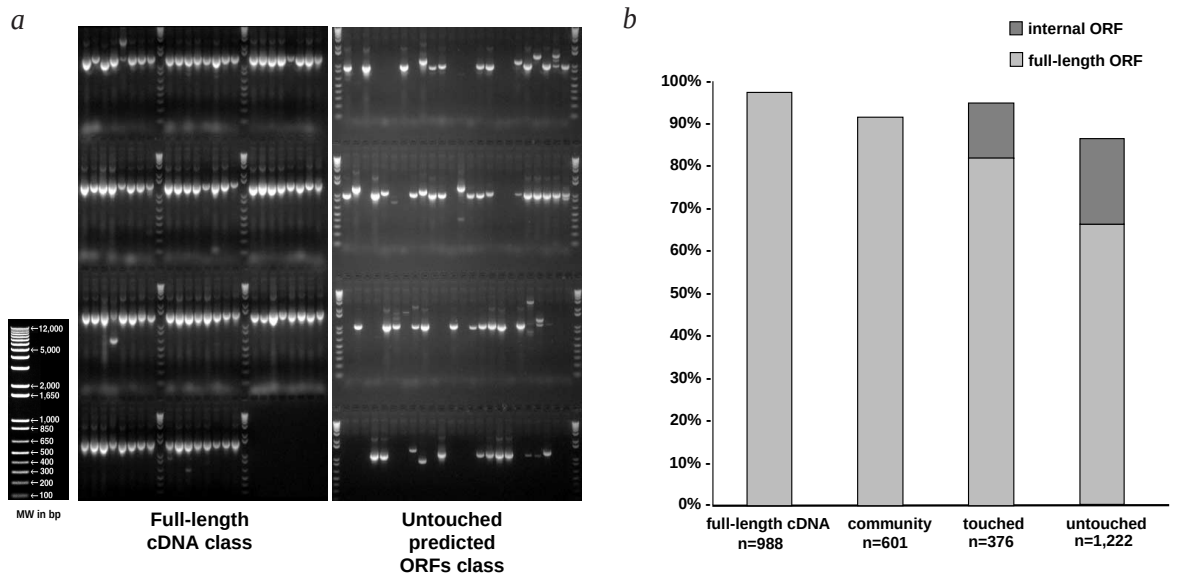


Fig. 2 PCR amplification of *C. elegans* ORFs. ***a***, Samples of PCR products obtained using primers that anneal just downstream of the ATG and just upstream of the stop codon as a way to amplify full-length protein-coding ORFs. Left, ORFs correspond to the 'full-length cDNA class' genes that were completely sequenced by the EST project. Right, ORFs correspond to completely unconfirmed or 'untouched' genes. A more complete set of data is available, including the pictures of the PCR products for the 'community' ORFs and the 'touched' ORFs (Web Fig. A), the pictures of the PCR products obtained from the internal primers (Web Fig. B), an analysis of the PCR success rates in relation to ORF sizes (Web Fig. C) and more detailed numbers (Web Table A). ***b***, Successful PCR reactions for the four classes of ORFs analyzed. The number of genes examined for each class is indicated below the class names. The percentages of PCR success shown for the touched and untouched classes is the combined total of successful PCR reactions using primers that amplified the full-length ORF and primers that amplified an internal portion of the ORF.

Thus 9,888 (18,959–9,071) predicted genes remain without experimental confirmation: the 'untouched' genes (Fig. 1). Because these predictions have not been confirmed by any ESTs, it is formally possible that a proportion represent GeneFinder false positives. Alternatively, these untouched genes might have remained undetected by the EST approach because they belong to a class of relatively weakly expressed genes[7]. We reasoned that the ORFeome cloning strategy should be more sensitive than the EST approach because it is based on the PCR amplification of ORFs (that is, directly from cDNA libraries and using specific primers (http://worfdb.dfci.harvard.edu)), as opposed to the random picking of cDNA clones.

We first tested our primer-design program and PCR conditions on previously identified genes that are relatively highly expressed and for which a full-length transcript sequence has been obtained (Y.K. *et al.*, manuscript in preparation). Of these 'full-length' genes, 97% gave rise to a PCR product of the expected size (Fig. 2). We then tested the quality of our cDNA library[4] using the community genes that, for most of them, were identified genetically by positional cloning. As such, they should represent a wide range of relative expression levels. Supporting this idea, 147 of 784 community genes were not present in the *C. elegans* EST database (Fig. 1). Of the community genes tested, 91% gave rise to a PCR product of the expected size (Fig. 2). Lastly, we wanted to determine our ability to amplify predicted ORFs for which the sequence surrounding the ATG and stop codons has not been confirmed, but for which one or a few ESTs have been identified so far (the 'touched' genes). We reasoned that GeneFinder predictions can be correct in predicting the existence of a gene, but incorrect in predicting either the 5′ or the 3′ end, or both. From a random sample of these touched genes, 83% gave rise to a PCR product, most of them of the expected size (Fig. 2).

Having established and tested a set of optimal conditions, we then applied the OST approach to a random sample representing approximately 12% of the approximately 10,000 predicted but unverified *C. elegans* genes. Approximately 66% gave rise to a PCR product (Fig. 2). The percentage of PCR products showing a size different than expected was significantly higher for the untouched genes (Fig. 3). It is expected that, among the touched and untouched genes that were not amplified (17% and 34%, respectively), at least a subset could be explained by inaccurate ATG and/or stop codon predictions. Hence, we attempted to detect these genes using primers that anneal within the predicted ORFs. For practical reasons we arbitrarily designed primers to amplify a PCR product of 300 bp. Because of this selected size, only 15% of touched genes and 26% of untouched genes could be tested (Fig. 2). This 'internal primer' experiment allowed the recovery of 83% of touched genes and 75% of untouched genes that failed to amplify in the initial PCR reactions using the ATG-stop primers (Fig. 2). Hence, an additional 13% (83%×15%) of touched genes and 20% (75%×26%) of untouched genes were detected in this way (Fig. 2).

Sequencing reactions were then carried out on the cloned ORFs from touched and untouched genes to generate OSTs. From alignments against the genomic sequence, most OSTs identified the ORF expected from the GeneFinder predictions and most revealed at least one splicing event. The OST analysis demonstrates that at least 93% of amplified touched genes and 86% of amplified untouched genes are correct and spliced (Table 1). In addition, 85% of touched and 66% of untouched internal products are correct and spliced as well (Fig. 2). Therefore, 88% ((93%×83%)+(85%×13%)) of our sample of 376 touched genes and at least 70% ((86%×66%)+(66%×20%)) of our sample of 1,222 predicted untouched genes could be experimentally confirmed using the OST approach (Fig. 2).

We then used this experimental ratio to improve the gene number estimate for *C. elegans*, in combination with the following parameters. First, we calculated the rate of false negatives of the OST approach. Using genes already demonstrated to exist but for which the ATG-stop surrounding sequences are not confirmed (that is,

the touched genes), we observed 12% failures of detection, for 88% success. Thus, assuming that the same rate of false-negative predictions can be applied to the untouched genes for which the ATG-stop surrounding sequences are also unconfirmed, the 70% of untouched genes that we detected should be prorated to 80% (70%×(1+(12%/88%))). We can thus conclude that 7,910 (9,888×0.80) untouched genes are likely to correspond to genuine genes. As stated previously, however, GeneFinder has a tendency to split genes (that is, to predict two genes for one existing gene). In some cases our approach would amplify both predicted genes even though the stop codon of the first one and the ATG of the second one are wrongly predicted. To estimate the ratio of such split predictions, we compared a set of sequenced full-length cDNAs with different versions of ACeDB GeneFinder predictions. We observed that the maximal ratio of genes split by GeneFinder is 8.4% (Table 2). We can thus estimate that at least 7,245 (7,910–(7,910×0.084))

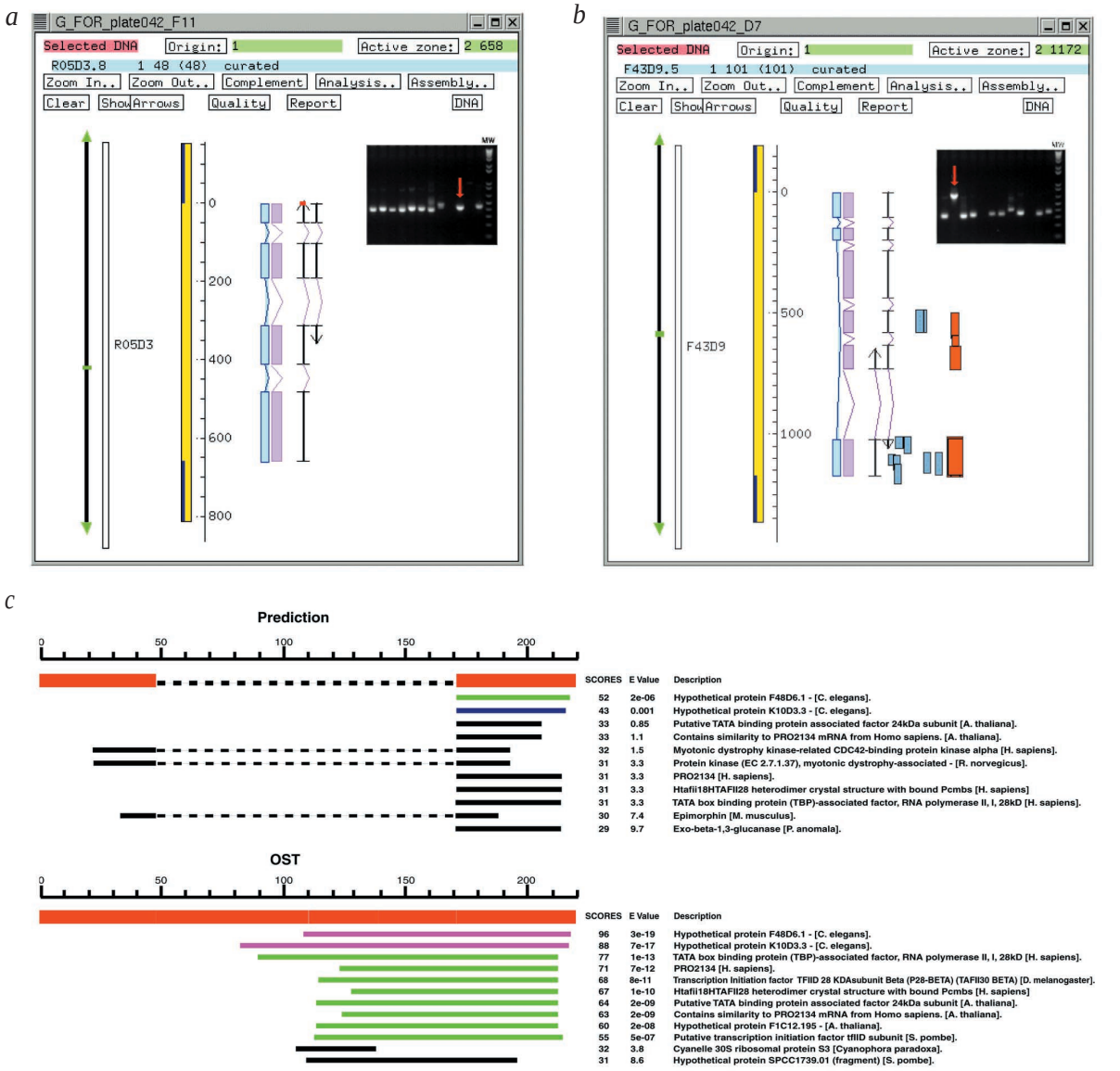


Fig. 3 Gene annotation using ORF sequence tags (OSTs). *a*,*b*, The Acembly output of the alignment for the predicted ORFs R05D3.8 and F43D9.5, respectively. In both panels, the continuous vertical bars (left) represent the *C. elegans* genome next to a number scale centered on the putative ATG for each ORF. The blue boxes correspond to GeneFinder predictions for exons, whereas the connecting blue lines indicate predicted introns. The black arrows (right) correspond to the OST sequencing reads. The OSTs are combined into observed exons (pink boxes) and introns (connecting pink lines). Insets, the agarose gel of the PCR products. ***a***, The observed PCR product (small arrow) is of the expected size and the sequencing indicates that the R05D3.8 OSTs match the predicted ORF for both size and structure, including exon/intron boundaries. ***b***, The observed PCR product (small arrow) is larger than expected. Nevertheless, sequencing did confirm that it corresponds to the F43D9.5 ORF, as it aligns with the predicted sequence. The exon/intron structure, however, does not completely match the GeneFinder prediction. In this case, GeneFinder failed to accurately identify three exons in F43D9.5. BlastX alignments with F43D9.5 as a query identify ORFs from *C. elegans* and *Homo sapiens* (filled blue boxes) that show similarity to the exons identified in the OSTs. TBlastX searches against *Caenorhabditis briggsae* (orange boxes) show concordant ORF and exon matches with the OSTs. ***c***, A representation of Blastp hits is shown using either the GeneFinder prediction for F43D9.5 (top alignment) or the OST sequence against the NCBI non-redundant protein database. The quality of the individual alignments are color-coded: pink, the highest alignment score (80–200); green, somewhat lower (50–80); blue, lower still (40–50), and black, insignificant (<40). The protein predicted from OSTs shows significant homologies with various transcription factors from 4 different species, 10 alignments with scores >50, whereas the protein predicted from the GeneFinder prediction results in scores too low to be considered significant.

Table 1 • OST analysis

	OSTs analyzed	Expected ORFs	Expected ORFs/ OSTs analyzed	Expected ORFs spliced	Exp ORFs-spliced/ OSTs analyzed
Touched	214	211	99%	199	93%
Untouched	655	617	94%	564	86%

The number of OSTs analyzed is indicated. Expected ORFs correspond to OSTs that demonstrate a correct identity. Expected ORFs spliced correspond to those ORFs for which we could detect at least one splicing event. More details can be obtained from Web Table A.

untouched genes correspond to genuine genes. When adding this number to the number of identified genes that were predicted by GeneFinder (8,707; Fig. 1), we obtained 15,952 as an estimate for the gene number based on GeneFinder predictions. Finally, this number should be corrected by the 9% rate of false negatives for the ability of GeneFinder to detect genes (796 genes not predicted for 8,707 identified; Fig. 1). In conclusion, the gene number in *C. elegans* can now be estimated to be at least 17,387 (15,952×1.09).

In addition to verifying the existence of untouched predicted genes, the OST approach allows the verification or correction of many intron/exon boundaries (Fig. 3). For most ORFs that gave rise to a PCR product of the expected size, the intron/exon boundaries observed in the OSTs corresponded exactly to those predicted by GeneFinder (Fig. 3*a*). Conversely, in many cases in which the observed size of an ORF's PCR product was different from the predicted size, OSTs were used to correct the GeneFinder predictions (Fig. 3*b*). Overall, 12% of the exons that were sequenced could be corrected (Fig. 4). These corrections correspond to 27% of the genes analyzed (151/564). In a few cases, such corrections identified new orthologies between predicted *C. elegans* proteins and proteins of known function in other organisms (Fig. 3*c*).

The predicted number of genes in *C. elegans*[1] was challenged by the annotation of the *Drosophila* genome leading to the prediction of 13,601 genes[2]. One possibility was that the *ab initio* GeneFinder predictions for *C. elegans* were too relaxed, introducing a large number of false positives. The work presented here largely excludes this possibility. Another plausible explanation was that the gene-prediction programs used for *Drosophila* are too stringent, perhaps because they are mostly based on known genes and ESTs. An accompanying report that describes an alternative *ab initio* genome annotation tool indicates that this may be the case in *Drosophila*[8]. Thus, *Drosophila* and *C. elegans* may have approximately the same number of genes. If so, it will be interesting to identify the determinants of the higher biological complexity of *Drosophila*.

Finally, our findings might shed light on the current debate for the estimates of the human gene number, which vary between 28,000 and 81,000 (refs. 9–12). Lower estimates usually rely on genome annotation tools that depend on ESTs or other experimental evidence. Our work, however, indicates that substantial numbers of biologically relevant human genes might not be detectable by ESTs and/or expression profiling approaches, due to their relatively low levels of expression. We suggest that the OST approach described here is an alternative strategy for gene verification of *ab initio* predictions that reduces the dependency on expression levels. In theory the OST approach should be applicable to human gene verifications provided that improved annotations can be derived from the use of refined algorithms and/or comparisons of the human genome with the genomes of other mammals.

Our work provides experimental evidence both for the number of genes in *C. elegans* and for the predicted proteome. In addition, we now have evidence that a comprehensive ORFeome cloning project is feasible for *C. elegans*. We can clone up to 97% of the ORFs that correspond to the approximately 5,000 genes for which a full-length cDNA sequence is currently available (Y.K., J.T.-M. and D.T.-M., manuscript in preparation), approximately 80% of the touched ORFs and approximately 65% of the ORFs that correspond to untouched genes. It is important, however, to note that only a subset of the latter class will correspond to correct ATG-to-stop codon ORFs. This is due to the fact that the ORFeome cloning technique cannot detect 5′ and/or 3′ ORF extensions in actual genes compared with GeneFinder predictions. Altogether this suggests that, for *C. elegans* and probably for other organisms as well, ORFeome cloning projects will depend on extensive transcriptome sequencing projects.

Note: supplementary information is available on the Nature Genetics web site (http://genetics.nature.com/supplementary_info/).

Methods

Working definition for the concept of genes. The genes analyzed here were arbitrarily defined as genomic sequences that can be transcribed into protein-encoding transcripts. RNA-encoding genes, although obviously important, were excluded because our experimental approach fails to confirm them. Various transcripts were considered to derive from a single gene if they shared at least one exon, or part of an exon, in their coding sequence and/or 5′ or 3′ UTRs. Such transcripts can encode different proteins that usually, although not necessarily, share one or more amino-acid domains. We did not impose any minimal size for the ORFs of predicted genes. The smallest ORF tested here contains 69 nucleotides.

Table 2 • Analysis of the rate of split predictions from GeneFinder

	Full-length cDNAs	Corresponding GeneFinder matches	Observed	False positives	False positive rate
	1	1	2,675	0	
July	1	2	172	172	
1996	1	3	20	40	
(WS1)	1	4	1	3	
Total			2,868	215	7.50%
	1	1	4,412	0	
December	1	2	300	300	
1998	1	3	34	68	
(WS6)	1	4	6	18	
	1	5	3	12	
Total			4,755	398	8.40%
	1	1	4,956	0	
January	1	2	241	241	
2001	1	3	21	42	
(WS29)	1	4	5	15	
	1	5	1	4	
Total			5,224	302	5.80%

Data representing the analysis of three different versions of ACeDB are shown. We compared 5,386 full-length cDNA sequences with GeneFinder predictions using Acembly. The number of full-length cDNAs that match 1, 2, 3, 4 or 5 GeneFinder predictions is shown. The false-positive rate due to split predictions is calculated as: (number of observed events)×(GeneFinder matches–1). A similar analysis for "joined" predictions gave rise to a false-negative rate of 4.4% (WS1), 3.4% (WS6) and 3.1% (WS29) (data not shown).

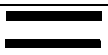
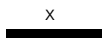
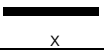
	exon unaltered	exon extended	exon shortened	additional intron	intron not found	additional exon	exon not found
geneFinder / OST							
observed events	1764	67	64	15	35	29	30
% of events	88.00%	3.34%	3.20%	0.75%	1.75%	1.45%	1.50%

Fig. 4 Alignments between GeneFinder predictions and OSTs. OSTs were aligned against the genome sequence and compared with GeneFinder predictions. We considered the following possibilities. The OST exons can be identical to GeneFinder-predicted exons (exon unaltered). The OST exons can be of different length compared with GeneFinder-predicted exons (they can be extended (exon extended) or shortened (exon shortened)). The OST exons can also be of a different structure compared with GeneFinder-predicted exons. There can be an additional intron (additional intron) or a missing intron (intron not found). Finally the OST approach can identify exons that were completely missed by GeneFinder (additional exon) or suggest that GeneFinder-predicted exons do not exist (exon not found). Numbers of events observed for each class are indicated along with the percentage of ORFs corrected.

Selection of random sets of genes. For the untouched genes, we first amplified all untouched genes of chromosome III (846) in the context of the ORFeome cloning project. To verify that there is no chromosome-specific bias, we selected 94 untouched genes from chromosomes I, II, IV and V using the following method. We first ordered the ORF names alphabetically. (ORF names derive from their cosmid or YAC designation followed by a letter or a number. The cosmid and YAC designations were assigned chronologically during the construction of the genome physical map. As such they should not be biased.) A set of ORFs was then compiled from the alphabetical list by picking every "x" ORF (with x=total number of ORFs in the list/number of ORFs desired). As expected, the OST analysis shows no significant differences among different chromosomes. For the touched genes, all chromosome III ORFs that have at least 1 EST were ordered alphabetically and the first 376 names were selected.

PCR amplification and Gateway cloning. ORFs corresponding to predicted genes were PCR amplified using primers directly downstream of the putative ATG and directly upstream of the corresponding in-frame STOP codon (ATG-STOP primers; J.R. *et al.*, manuscript in preparation) and a highly representative cDNA library as template DNA (ref. 4). PCR was carried out in 96-well format using a high-fidelity *Taq* DNA polymerase (Life Technologies). The cycling conditions were 94 °C for 2 min, 94° C for 45 s, 56 °C for 1 min, 68 °C for 1 min per kb for 35 cycles, and 68 °C for 5 min. After PCR, samples were analyzed on an agarose gel. We organized the experiments so that the predicted ORF sizes are arranged by increasing size. This way, the elongation time can be optimized for each plate individually. Discrepancies between expected and observed size for any PCR product are readily detected on visual inspection of the gels. For the amplification of internal segments of ORFs, the reverse primers were chosen arbitrarily 100 bp upstream of the putative STOP codon and the forward primers were systematically selected 300 bp upstream of the reverse primers and at least 100 bp downstream of the putative ATG. Consequently, ORFs smaller than 500 bp were not included in this analysis.

PCR products were cloned by *in vitro* recombination using the Gateway system[4–6]. The resulting recombinant molecules were transformed in DH5α and selected on plates containing kanamycin (50 μg/ml). Plasmid DNA was prepared from pools of transformants and subsequently sequenced.

OST alignment. Alignment of OST sequences against the *C. elegans* genomic sequence was carried out using Acembly on version WS9 of ACeDB. Each OST sequence was aligned against the complete *C. elegans* genomic sequence to confirm the identity of the individual OSTs and to compare the experimentally determined sequence with that found in ACeDB. Individual alignments were also checked for exon/intron boundaries to identify splicing events.

Acknowledgments

We thank S. Boulton, L. Matthews, J. Polanowska, M. Tewari and A.J.M. Walhout for comments on the manuscript and discussions; and L. Hillier and P. Green for the primer design program (OSP). This work was supported by grants from CREST, Japan Science and Technology Corporation and Grant-in-Aid for Scientific Research on Priority Areas from the Ministry of Education, Science, Sports and Culture of Japan (to Y.K.), and by grants 1 RO1 HG01715-01 from the National Human Genome Research Institute, 1 R21 CA81658 A 01 from the National Cancer Institute and 128 from the Merck Genome Research Institute (to M.V.).

Received 28 November 2000; accepted 6 February 2001.

1. The *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012–2018 (1998).
2. Adams, M.D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195 (2000).
3. The *Arabidopsis* Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000).
4. Walhout, A.J.M. *et al.* Gateway recombinational cloning: application to the cloning of large numbers of open reading frames, or ORFeomes. *Methods Enzymol.* **328**, 575–592 (2000).
5. Walhout, A.J.M. *et al.* Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science* **287**, 116–122 (2000).
6. Hartley, J.L., Temple, F.T. & Brasch, M.A. DNA cloning using *in vitro* site-specific recombination. *Genome Res.* **10**, 1788–1795 (2000).
7. Hill, A.A., Hunter, C.P., Tsung, B.T., Tucker-Kellogg, G. & Brown, E.L. Genomic analysis of gene expression in *C. elegans*. *Science* **290**, 809–812 (2000).
8. Gopal, S. *et al.* Homology-based annotation yields 1,042 new candidate genes in the *Drosophila melanogaster* genome. *Nature Genet.* **27**, 337–340 (2001).
9. Ewing, B. & Green, P. Analysis of expressed sequence tags indicates 35,000 human genes. *Nature Genet.* **25**, 232–234 (2000).
10. Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **402**, 489–495 (1999).
11. Roest Crollius, H. *et al.* Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. *Nature Genet.* **25**, 235–238 (2000).
12. Liang, F. *et al.* Gene Index analysis of the human genome estimates approximately 120,000 genes. *Nature Genet.* **25**, 239–240 (2000); correction: **26**, 501 (2000).